A Speaker Independent Continuous Speech Recognizer for Amharic

Hussien Seid

Computer Science & Information Technology
Arba Minch University
PO Box 21, Arba Minch, Ethiopia
huss1438@yahoo.com

Björn Gambäck

Userware Laboratory
Swedish Institute of Computer Science AB
Box 1263, SE-164 29 Kista, Sweden
gamback@sics.se

Abstract

The paper discusses an Amharic speaker independent continuous speech recognizer based on an HMM/ANN hybrid approach. The model was constructed at a context dependent phone part sub-word level with the help of the CSLU Toolkit. A promising result of 74.28% word and 39.70% sentence recognition rate was achieved. These are the best figures reported so far for speech recognition for the Amharic language.

1. Introduction

The general objective of the present research was to examine and demonstrate the performance of a hybrid HMM/ANN system for a speaker independent continuous Amharic speech recognizer. Amharic is the official language of communication for the federal government of Ethiopia and is today probably the second largest language in the country (after Oromo) and quite possibly one of the five largest on the African continent. It is estimated to be mother tongue of more than 17 million people, with at least an additional 5 millions of second language speakers. Still, just as for many other African languages, Amharic has received preciously little attention by the speech processing research community; even though the last years have seen an increasing trend to investigate applying speech technology to other languages than English, most of the work is still done on very few and mainly European and East-Asian languages.

The Ethiopian culture is ancient, and so are the written languages of the area, with Amharic using its very own script. This has caused some problems in the digital age and even though there are several computer fonts for Amharic, and an encoding of Amharic was incorporated into Unicode in 2000, the language still has no widely accepted computer representation. In recent years there has been an increasing awareness of that Amharic speech and language processing resources must be created as well as digital information access and storage.

The present paper is a step in that direction. It is laid out as follows: Section 2 introduces the HMM/ANN hybrid ASR paradigm. Section 3 discusses various aspects of Amharic and some previous efforts to apply speech technology to the language. Then Section 4 describes the actual experiments with constructing, evaluating, and testing an Amharic Automatic Speech Recognition System using the CSLU Toolkit [1].

2. HMM/ANN hybrids

Commonly, HMM-based speech recognizers have shown the best performance. On the positive side this dominant paradigm is based on a rich mathematical framework which allows for powerful learning and decoding methods. In particular, HMMs are excellent at treating temporal aspects by providing good abstractions for sequences and a flexible topology for statistical phonology and syntax. However, HMMs have some drawbacks, especially for large vocabulary speaker independent continuous ASR. The main disadvantage is a relatively poor discrimination power. In addition HMMs enforce some practical requirements for distributional assumptions (e.g., uncorrelated features within an acoustic vector) and typically make first order Markov model assumptions for phone or sub-phone states while ignoring the correlation between acoustic vectors [2].

In effect, HMMs adopt a hierarchical scheme modeling a sentence as a sequence of words, and each word as a sequence of sub-word units. An HMM can be defined as a stochastic finite state automaton, usually with a left-to-right topology when used for speech. Each probability is approximated based on maximum likelihood techniques. Still, these techniques have been observed for poor discrimination, since they maximize the likelihood of each individual node independently from the other. On the other hand neural network classifiers have shown good discrimination power, typically requires fewer assumptions, and can easily be integrated in non-adaptive architectures. This is the point behind changing the pure HMM approach to the hybrid HMM/ANN model, by using an ANN to augment the ASR system [3]. The HMM is used as the main structure of the system to cope with the temporal alignment properties of the Viterbi algorithm, while the ANN is used in a specific subsystem of the recognizer to address static classification tasks. This has shown performance improvement over pure HMM: Fritsch & Finke [4] describe a tree-structural hierarchical HMM/ANN system which outperformed HMM on Switchboard.

In an HMM/ANN model a neural network of multi-layered perceptrons is given an input vector of acoustic observation values, o_t and computes a vector of output values which are approximate a-posteriori state probabilities. Commonly, nine frames are given for the input of the network: four consecutive frames before, four frames after, and one frame at time t, in order to provide the ANN with more contextual data. Then the network will have one output for each phone by restricting the sum of all the output units to one. This helps to calculate the a-posteriori probability, q_j of a state j conditioned on the acoustic input: $p(q_j|o_t)$. Generally an ASR system has a front end in which the natural speech wave is digitized and parameterized for the recognizer. The recognizer has a neural net to train on these digitized and parameterized data. After training, the neural net produces the estimation of probabilities of observations for the HMM states. The HMM uses these probabilities and the language model to compute the probability of a sequence of symbols given the observation sequence. Finally, the recognizer uses decoders to generate the recognized symbols as output.

3. Amharic Speech Processing

Ethiopia is with about 70 million inhabitants the third most populous African country and harbours some 80 different languages. Three of these are dominant: Oromo, a Cushitic language is spoken in the South and Central parts of the country and written using the Latin alphabet; Tigrinya, spoken in the North and in neighbouring Eritrea; and Amharic, spoken in most parts of the country, but predominantly in the Eastern, Western, and Central regions. Amharic and Tigrinya are Semitic languages and thus distantly related to Arabic and Hebrew.

3.1. The Amharic language

Following the Constitution of 1994, Ethiopia is a divided into nine fairly independent regions, each with its own nationality language. However, Amharic is the language for country-wide communication and was also for a long period the principal language for literature and the medium of instruction in primary and secondary schools of the country (while higher education is carried out in English). Amharic speakers are mainly Orthodox Christians, with Amharic and Tigrinya drawing common roots to the ecclesiastic Ge'ez still used by the Coptic church — both languages are written horizontally and left-to-right using the Ge'ez script. Written Ge'ez can be traced back to at least the 4th century A.D. The first versions of the language included consonants only, while the characters in later versions represent consonant-vowel (CV) phoneme pairs.

Amharic words use consonantal roots with vowel variation expressing difference in interpretation. In modern written Amharic, each syllable pattern comes in seven different forms (called *orders*), reflecting the seven vowel sounds. The first order is the basic form; the other orders are derived from it by more or less regular modifications indicating the different vowels. There are 33 basic forms, giving $7 * 33$ syllable patterns (syllographs), or *fidEls*. Two of the base forms represent vowels in isolation (ዐ and አ), but the rest are for consonants (or semivowels classed as consonants) and thus correspond to CV pairs, with the first order being the base symbol with no explicit vowel indicator (though a vowel is pronounced: C+/ə/). The writing system also includes four (incomplete, five-character) orders of labialised velars and 24 additional labialised consonants. In total, there are 275 *fidEls*. See, e.g., [5] for an introduction to the Ethiopian writing system.

The Amharic writing system uses multitudes of ways to denote compound words and there is no agreed upon spelling standard for compounds. As a result of this — and of the size of the country leading to vast dialectal dispersion — lexical variation and homophony is very common. In addition, not all the letters of the Amharic script are strictly necessary for the pronunciation patterns of the spoken language; some were simply inherited from Ge'ez without having any semantic or phonetic distinction in modern Amharic. There are many cases where numerous symbols are used to denote a single phoneme, as well as words that have extremely different orthographic form and slightly distinct phonetics, but with the same meaning. So are, for example, most labialised consonants basically redundant, and there are actually only 39 context-independent phonemes (monophones): of the 275 symbols of the script, only about 233 remain if the redundant ones are removed.

In contrast to the character redundancy, there is no mechanism in the Amharic writing system to mark gemination of consonants. The words /wɛnɛ/ (swimming) and /wɛnnɛ/ (main, core) are both written as ዋና, but give two completely different meanings by geminating the consonant ን /n/. This requires different reference models in the database for the multiple forms of the sound depending on the gemination. (Another problem is an ambiguity with the 6th order characters: whether they are vowelled or not. However, this is not relevant to this work.)

3.2. Previous work

This study aims at investigating and testing out the possibility of developing speaker independent continuous Amharic speech recognition systems using a hybrid of HMM and ANN systems. Speech and language technology for the languages of Ethiopia is still very much unchartered territory; however, on the language processing side some initial work has been carried out, mainly on Amharic word formation and information access. See [6] or [7] for short overviews of the efforts that have been made so far to develop language processing tools for Amharic.

Research conducted on speech technology for Ethiopian languages has been even more limited. Laine [8] made a valuable effort to develop an Amharic text-to-speech synthesis system, and Tesfay [9] did similar work for Tigrinya.[1] Solomon [10] built speaker dependent and speaker independent HMM-based isolated consonant-vowel syllable recognition systems for Amharic. He proposed that CV-syllables would be the best candidates for the basic recognition units for Amharic.

Solomon's work was extended by Kinfe [11] who used the HTK Toolkit to build HMM word recognizers at three different sub-word levels: phoneme, tied-state triphone, and CV-syllable. Kinfe collected a 170 word vocabulary from 20 speakers. He considered a subset of the Amharic syllables, concentrating on the combination of 20 phonemes with the seven vowels, or in total 140 CV-units. Kinfe's training and test sets both consisted of 50 discrete words. Contrary to Solomon's predictions, the performance of the syllable-level recognition was very bad (for unclear reasons) and Kinfe abandoned it in favour of the phoneme- and triphone-based recognizers. For the latter two he reports an isolated word recognition accuracy of 83.1% resp. 78.0% on speaker dependent models, while the speaker independent models gave 75.5% for phoneme-based models and 77.9% isolated word accuracy for tied-state triphone models.

Molalgne [12] tried to compare HMM-based small vocabulary speaker-specific continuous speech recognizers built using three different toolkits: CSLU, HTK, and MSSTATE Toolkit from Mississippi State, but failed in setting up CSLU so that only two toolkits were actually tested. He collected a corpus of 50 sentences with ten words (the digits) from a single speaker. While HTK was clearly faster than MSSTATE, the speaker dependent recognition performance for both systems was comparable with 82.5% resp. 79.0% word accuracy and 72.5% resp. 67.5% sentence accuracy for HTK resp. MSSTATE.

Martha [13] worked on a small vocabulary isolated word recognizer for a command and control interface to Microsoft Word, while Zegaye [14] continued the work on speaker independent continuous Amharic ASR. He used a pure HMM-based approach and reached 76.2% word accuracy and 26.1% sentence level accuracy. However, there are still a lot of work to be done towards achieving a full-fledged automatic Amharic speech recognition system. The intention of the present research was to use an HMM/ANN hybrid model approach as an alternative for better performance. For this we utilized an implementation of such a model in the CSLU Toolkit.

[1] In the text we follow the practice of referring to Ethiopians by their given names. However, the reference list follows European standard and also gives surnames (i.e., the father's given name for an Ethiopian).

4. An Amharic SR system

The attempt of this research is to design a prototype speech recognizer for the Amharic language. The recognizer uses phonemes as base units and is designed to recognize continuous speech and is speaker independent. In contrast to the pure HMM-based work done by Zegaye [14], the system implements the HMM/ANN hybrid model approach. The development process was performed using the CSLU Toolkit installed on the Microsoft Windows 2000 platform. Various preprocessing programs and script editors were used to handle vocabulary files.

4.1. The CSLU Toolkit

The CSLU Toolkit [1] was designed not only for speech recognition, but also for research and educational purposes in the area of speech and human-computer interactions. It is developed and maintained by the Center of Speech Language Understanding, a research centre at the Oregon Graduate Institute of Science and Technology, Portland and the Center for Spoken Language Research at the University of Colorado. The toolkit, which is available free of charge for educational, research, personal, and evaluation purposes under a license agreement, supports core technologies for speech recognition and speech synthesis, plus a graphical based rapid application development environment for developing spoken dialogue systems.

The toolkit supports the development of HMM or HMM/ANN hybrid-based speech recognition systems. For this purpose it has many modules or tools interacting with each other in an environment called CSLU-HMM. The toolkit needs a consistent organization and naming of directories and files which has to be strictly followed. This is tedious work, but also clearly doable (still, this might have been the reason why Molalgne decided that it was not possible to use the CSLU Toolkit [12]).

4.2. Speech data

Apart from the specifics of the language itself, the main problem with doing speech recognition for an under-resourced language like Amharic is the lack of previously available data: No standard speech corpus has been developed for Amharic. However, we were able to use a corpus of 50 speakers recorded at 16 kHz sampling rate by Solomon [10]. 100 different sentences of read speech were recorded for each speaker.

The corpus was prepared and processed using `SpeechView`, a part of the CSLU Toolkit providing a graphic-based interface to prepare speech data. The tool is used to record, display, save, and edit speech signals in their wave format. It also provides spectrograms and some other speech wave related data like pitch and energy counters, neural net outputs, and phonetic labels. With the help of the `SpeechView` tool, one can collect and prepare speech data in an easy way for training a recognizer. The process of annotating the speech waveform, which is the most tedious and difficult process in the development of speech recognition systems, can be done at different transcription levels.

Ten spoken sentences each from ten female speakers were annotated at the phoneme level for the training corpus and time-aligned word level transcriptions were generated automatically. Two more speakers were annotated for evaluation purposes. Long silences at the beginning and end of the wave file were trimmed off and the boundaries of word-level transcriptions were adjusted accordingly.

A vocabulary file was created based on the pronunciation of each word in the data set and parts of the phones. This gave a vocabulary of 778 words represented by 34 phones that in turn were split into 57 phone parts: ቭ, ኽ, ጵ, and ዥ were defined to consist of three parts each; 15 phones have two parts (ኣ, ም, ስ, ግ, ክ, ኤ, ቅ, ጥ, ፍ, ዝ, ኝ, ፕ, ዕ, ጅ, and ጭ), while 15 have one part only (እ, ን, ት, ይ, ል, ሙ, ር, አ, ሽ, ብ, ድ, ኢ, ኡ, ሀ, and ቭ). Each phone group is here ordered internally according to frequency.

4.3. Experiments

Thereafter a recognizer was created, the frame vectors were generated automatically in the toolkit, and the recognizers was trained on the phone part files. The ANN of the recognizer contained an output layer with the phone parts, while the input layer was a 180 node grid representing 20 features each from nine time frames ($t \pm 4 * 10$ms).

The recognizer was evaluated on two sentences each from ten speakers who were all found in the training data (in total 20 sentences and 236 words). The results were as shown in Table 1.

Itr	Subst	Insert	Delete	Word Acc	Snt Corr
15	13.62	4.89	5.83	75.66	42.31
16	13.62	5.83	5.83	74.72	42.31
17	13.62	4.89	6.83	74.67	41.72
18	14.61	4.89	5.83	74.67	42.31
19	15.56	3.89	4.89	75.66	41.72
20	11.67	5.79	4.89	77.65	42.90
21	11.67	5.83	4.89	77.61	42.90
22	14.61	5.83	5.83	73.73	41.13
23	13.62	4.89	4.89	76.61	42.90
24	13.62	2.93	5.79	77.66	42.90
25	14.61	2.93	4.89	77.57	42.31
26	14.61	4.89	4.89	75.62	42.31
27	15.56	3.89	4.89	75.66	42.31
28	12.66	3.89	4.89	**78.56**	**44.07**
29	12.66	5.83	4.89	76.62	42.31
30	12.66	4.89	4.89	77.56	42.90

Table 1: Recognition accuracy on known speakers.
Best result: 78.56% word and 44.07% sentence level accuracy.

Itr	Subst	Insert	Delete	Word Acc	Snt Corr
15	16.34	5.87	7.00	70.79	35.27
16	16.34	7.00	7.00	69.65	35.17
17	16.34	5.87	8.20	69.59	33.79
18	17.53	5.87	7.00	69.60	34.27
19	18.68	4.66	5.87	70.80	33.79
20	14.00	6.93	5.87	73.20	36.75
21	14.00	7.00	5.87	73.13	35.35
22	17.53	7.00	7.00	68.46	33.62
23	16.34	5.87	5.87	71.92	37.75
24	16.34	3.52	6.95	73.19	34.75
25	17.53	3.52	5.87	73.08	34.27
26	17.53	5.87	5.87	70.73	34.27
27	18.68	4.66	5.87	70.80	34.27
28	15.19	4.66	5.87	**74.28**	**39.70**
29	15.19	7.00	5.87	71.94	35.27
30	15.19	5.87	5.87	73.07	35.64

Table 2: Recognition accuracy on unknown speakers.
Best result: 74.28% word and 39.70% sentence level accuracy.

For each iteration the columns in Table 1 give the percentage of substitutions, insertions, and deletions, as well as the word accuracy, and the percentage of correct sentences. The best results (78.56% word level accuracy and 44.07% sentence correctness) were obtained after 28 iterations.

When the same recognizer was tested for another ten speakers who were not included in the training data with two sentences each (218 words in total), the recognition rate degraded. As can be seen in Table 2, the best results were again obtained after the 28th iteration. The word accuracy was reduced by 4.28%, while the sentence level recognition rate was reduced by 4.37%, giving a 21.44% word level error rate and 55.93% sentence level error rate.

Accordingly, the HMM/ANN hybrid recognizer gave a 2.36% decrease in word error rate and 18.01% decrease in sentence error rate compared to Zegaye's purely HMM-based recognizer [14], which had 23.80% word and 73.94% sentence error rates. The relative error reduction compared to Zegaye's work is thus 9.92% at the word level and 24.36% at the sentence level.

5. Conclusions

The paper reported experiences with using the CSLU Toolkit to build a hybrid HMM/ANN speaker independent continuous speech recognizer for Amharic, the main language of Ethiopia. An annotated corpus was created from previously recorded speech data. Ten sentences each from twelve speakers were marked up at the phoneme level and a vocabulary of 778 words was created.

For speakers found in the training data, the best results obtained were 78.6% word and 44.1% sentence level accuracy. When tested on data from ten previously unseen speakers, the recognizer had a 74.3% word accuracy and 39.7% sentence accuracy; a relative error reduction of 24.4% compared to previous work on Amharic, using pure HMM-based methods.

The CSLU Toolkit proved to be a good vehicle to develop hybrid HMM/ANN-based recognizers, and the experiments indicate that a better recognizer can be developed with further optimization efforts. However, the implementation of the toolkit in Windows needs some revisions. There were problems to fully download the Toolkit Installer and after installation the system integration with Windows required considerable efforts.

6. Acknowledgements

This research was carried out at the Department of Information Science, Addis Ababa University and could not have come into being without the help of Solomon Berhanu who provided the corpus. Thanks to Zegaye Seifu and Kinfe Tadesse for constructive comments and to Marek F. and Clemente Fragoso Eduardo for help with fixing CSLU Toolkit implementation problems.

The work was funded by the Faculty of Informatics at Addis Ababa University and the ICT support programme of SAREC, the Department for Research Cooperation at Sida, the Swedish International Development Cooperation Agency.

7. References

[1] J.-P. Hosom, R. Cole, M. Fanty, J. Schalkwyk, Y. Yan, and W. Wei, "Training neural networks for speech recognition," Webpage, Feb. 1999. [Online]. Available: speech.bme.ogi.edu/tutordemos/nnet_training/tutorial.html

[2] H. Bourlard and N. Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," in *Adaptive Processing of Sequences and Data Structures*, C. Giles and M. Gori, Eds. Springer-Verlag, 1997, pp. 389–417.

[3] F. Beaufays, H. Bourlard, H. Franco, and N. Morgan, "Neural networks in automatic speech recognition," in *The Handbook of Brain Theory and Neural Networks*, 2nd ed., M. Arbib, Ed. MIT Press, 2002, pp. 1076–1080.

[4] J. Fritsch and M. Finke, "ACID/HNN clustering hiearchies of neural networks for context-dependent connectionist acoustic modeling," in *Proc. International Conference on Acoustics, Speech and Signal Processing*. Seattle, Washington: IEEE, Apr. 1998, pp. 505–508.

[5] T. Bloor, "The Ethiopic writing system: a profile," *Journal of the Simplified Spelling Society*, vol. 19, pp. 30–36, 1995.

[6] Atelach Alemu, L. Asker, and Mesfin Getachew, "Natural language processing for Amharic: Overview and suggestions for a way forward," in *Proc. 10th Conference 'Traitement Automatique des Langues Naturelles'*, vol. 2, Batz-sur-Mer, France, June 2003, pp. 173–182.

[7] Samuel Eyassu and B. Gambäck, "Classifying Amharic news text using Self-Organizing Maps," in *Proc. 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan, June 2005, Workshop on Computational Approaches to Semitic Languages.

[8] Laine Berhane, "Text-to-speech synthesis of the Amharic language," MSc Thesis, Faculty of Technology, Addis Ababa University, Ethiopia, 1998.

[9] Tesfay Yihdego, "Diphone based text-to-speech synthesis system for Tigrigna," MSc Thesis, Faculty of Informatics, Addis Ababa University, Ethiopia, 2004.

[10] Solomon Berhanu, "Isolated Amharic consonant-vowel syllable recognition: An experiment using the Hidden Markov Model," Msc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, 2001.

[11] Kinfe Tadesse, "Sub-word based Amharic speech recognizer: An experiment using Hidden Markov Model (HMM)," MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, June 2002.

[12] Molalgne Girmaw, "An automatic speech recognition system for Amharic," MSc Thesis, Dept. of Signals, Sensors and Systems, Royal Institute of Technology, Stockholm, Sweden, Apr. 2004.

[13] Martha Yifiru, "Automatic Amharic speech recognition system to command and control computers," MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, 2003.

[14] Zegaye Seifu, "HMM based large vocabulary, speaker independent, continuous Amharic speech recognizer," MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, 2003.